百度大数据应用与实践

陈尚义
百度公司 北京 100085

摘要
产生于互联网的大数据应用，现阶段正在向其他行业领域渗透，成为行业创新和转型的重要驱动力。根据百度多年来在大数据领域的创新与实践，阐述了大数据驱动搜索引擎的发展，介绍了百度大数据引擎和行业应用实践。重点分析了大数据发展的关键因素，并提出了大数据和人工智能是未来信息技术发展的重要方向。

关键词
大数据；人工智能；搜索引擎

Big Data Applications and Practices of Baidu

Chen Shangyi
Baidu.com.Inc., Beijing 100085, China

Abstract
Big data and the related applications which derived from the internet originally, are now expanding to other industries, and becoming the key driving force of their innovation and transition. The evolvement of the search engine driven by big data technologies was described, based on Baidu's innovations and practices in the big data area over the years. Baidu big data engine and its explorations in other industries were introduced. Finally, a vision was discussed that big data and artificial intelligence will be prospected in the future information communication technology.

Key words
big data, artificial intelligence, search engine

1 引言

随着移动互联网、物联网的快速发展，信息采集成本不断降低，加速物理世界向网络空间的量化。数字世界与现实世界的融合过程中产生并积累了大量的数据。根据国际数据公司（IDC）发布的研究报告，全球所有信息数据中90%产生于近几年，数据总量正在以指数形式增长，从2003年的5 EB[1]，到2013年4.4 ZB[2]，并将于2020年达到44 ZB，如图1所示。

数据爆炸将我们推向大数据时代，大数据是新一轮信息技术革命与人类经济社会活动的交汇融合的必然产物，数据的关联和挖掘将创造新的价值，提升效率。数据将和自然资源、人力资源一样成为国家最重要的战略资源，将成为产业升级的重要推动力。

大数据因其蕴含的社会价值和商业价值，已经成为一项重要的生产要素，大数据的应用将改变传统行业的商业模式，拉动产业升级。数据已经成为传统行业的核心资产。产生于互联网的大数据应用，现阶段正在向制造业、金融及商业、医疗卫生、国计民生等各个领域渗透。各行业也已经意识到数据价值挖掘的重要意义，加速探索并布局大数据应用。越来越多机构、企业都迫切希望从不同渠道获取的多种类型、结构复杂的大数据中挖掘出有价值的趋势洞察，快速、准确地制定决策，驱动商业和行业创新。

2 从搜索引擎说起，大数据面面观

2.1 搜索引擎是个天然的大数据服务

大数据是信息技术及其应用发展到一定阶段的“自然现象”，源于信息技术的不断廉价化以及互联网及其所带来的无处不在的信息技术延伸应用。可以说大数据应用和技术是在互联网的快速发展中产生的，互联网企业尤其是搜索引擎公司是大数据实践的先行者和领跑者。搜索引擎连接了人和信息、人和服务，本身就是一个完美的大数据应用实例，其目的就是为了更好地理解用户的搜索需求，将信息与用户匹配起来。

百度是当今中国人获取信息的最主要入口，每天响应来自138个国家和地区的数十亿次搜索请求，覆盖95%以上的中国网民，平均每个中国网民每天使用10次百度。为了获得更好的用户体验和搜索的精准对接，百度不断在技术上挑战自我，在搜索的

1 Exabyte（EB），1 EB=1 024 PB =2^{20} TB=2^{60} byte

2 Zettabyte（ZB），1 ZB=1 024 EB

3 来源于 IDC 报告

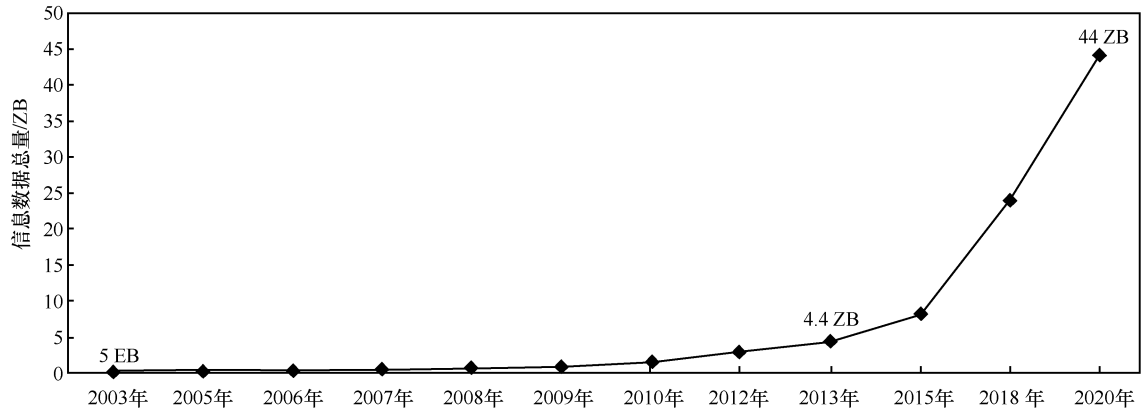


图1 全球数据总量[3]

实践中积累了整套大数据的处理和实践技术，占据了世界领先的地位。同时，百度也积极在大数据的商业实践上不断探索，并取得了显著的成绩。

2.2 海量的数据资源是大数据实践的基础

百度拥有海量的数据基础，拥有EB级别的超大数据存储和管理规模，并达到100 PB/天的数据计算能力，可达到毫秒级响应速度。百度已收录全世界超过一万亿张网页，相当于5 000个国家图书馆的信息量总和[4]。同时承担着每天百亿次的访问请求，可离线完成1 000亿网页的处理与分析，时效性网页从更新到索引只需要几十秒，实现大数据量级下的低延迟和秒级响应。

百度的数据具有实时性和全面性的特点，囊括了全网搜索数据、全网评论信息、百度内部数据以及第三方合作数据等跨行业、跨地域基础数据，海量的数据基础是百度引领大数据实践的基础。

2.3 高效的云计算基础设施提供强大的计算能力

面临庞大数据量带来的计算能力和网络带宽的新挑战，百度自主研发超大规模分布式存储和计算系统，目前能够支持14款用户过亿的产品[5]。其中分布式存储系统可以存储长文本、语音、视频等异构数据，实现单集群文件数达100亿；大规模分布式计算系统通过自研技术提升50%以上MapReduce的性能，实时流计算系统吞吐量达10 GB/s；百度创新性地实现了基于大数据的智能自动化运维框架，满足超大规模集群运维的需求，实时分析3万以上监控指标；2 min内完成分析和故障定位，保证系统可用性为99.99%。百度是全球首家大规模商用ARM服务器的公司，建立了大规模GPU并行化平台，单GPU计算能力可比百片CPU，极大程度地降低了能耗和计算成本。

百度自主研发万兆交换机，逐步从吉比特网络向万兆网络大规模切换，正在研制的4万兆交换机也已经开始小规模试点和验证，百度的万兆集群是国内互联网行业首个万兆交换机的规模应用。

基于完全自主知识产权的高性能服务器、整机柜和网络设备等，百度自主设计并建设了数个亚洲一流的数据中心，自主研发了整机柜服务器并已投入使用数十万台。通过基础设施、IT设备及软件协同，定制低功耗服务器等多项绿色节能技术，百度自建数据中心全年约一半时间实现完全免费冷却（free cooling）。2013年，该数据中心最佳PUE(power usage effectiveness，电力使用效率)为1.16，成为国内最节能、最环保的数据中心。

2.4 人工智能技术全面提升大数据处理能力

百度高度重视人工智能技术的发展，经过多年的坚持努力，在语音识别、图像识别、自然语言理解、机器学习、智能交互、数据挖掘、个性化推荐的研究和应用领域打下扎实的技术积累，攻克多项技术难题，人工智能技术已经达到国际领先水平。

百度目前已拥有全球最大规模的深度神经网络，并实现全球最大规模的GPU并行计算平台。百度的深度学习技术被应用在语音、图像、文本识别、自然语言处理和CTR预估等商业产品领域，取得显著的成效。同时，百度也积极将人工智能技术应用于大数据领域，通过机器学习和深度学习等技术提升数据智能，寻求现有问题的解决方案，并实现更好的预测。

4 http://tech.qq.com/a/20140529/023965.htm

5 http://www.china.com.cn/news/tech/2014-07/16/content_32972136.htm

3 大数据推动搜索引擎的演进

以百度为例，用户在搜索的过程中留下信息，其中有大量的文本、图片和影音等数据，形成了海量的数据资源，百度对这些复杂的异构数据进行处理分析，发掘价值，实现更多大数据应用。大数据技术推动着搜索引擎不断向前演进。

3.1 智能交互

随着用户需求更趋于复杂化和个性化，从最初的获取信息，到现阶段希望能够通过搜索引擎直接获取答案、连接服务，这就需要实现海量数据的挖掘和智能处理，实现人和服务的精准匹配。另外用户也更趋向于自然的交互方式，据统计，现阶段在百度的搜索请求中10%是以语音的形式表达的，而未来5年使用语音和图像来表达需求的比例将超过50%。基于如此真实强大的需求，为了不断提升用户体验，百度在图像识别和语音识别这两项前沿技术领域实现突破，并取得了一系列领先成果。

百度在2010年开始进行智能语音及相关技术研发，推出了第一代基于云端识别的互联网应用“掌上百度”。2012年11月，百度上线了中国第一款基于DNN的汉语语音搜索系统，成为最早采用DNN技术进行商业语音服务的公司之一。目前已经积累了数万小时的声学训练语料和海量文本语料[1]，线上语言模型体积超过100 GB，支持小时级别的海量语言模型更新。语音识别DNN深达9层，基于听觉感知的深度学习声学建模技术可以实现更高的精准度和识别率。在安静情况下，百度的普通话识别率已达到95%以上，处于国际领先水平。百度语音技术对内应用于手机百度、百度输入法、百度地图、百度导航等一系列产品，同时对外推出开放平台，提供多个垂直领域的识别和解析服务，合作伙伴超过30个，覆盖汽车、医疗、手机、电商、家电和车载等十几个领域和方向。

在图像识别领域，百度在2012年底将深度学习技术成功应用于OCR识别和人脸识别，并推出相应的PC端和移动端搜索产品[2]。2013年，深度学习模型被成功应用于一般图片的识别和理解。目前百度的人脸识别准确率超过98%，处于国际领先水平，图像识别技术已经用于手机百度、百度识图等多个应用中。从百度的经验来看，深度学习应用于图像识别不但大大提升了准确性，而且避免了人工特征抽取的时间消耗，从而大大提高了在线计算效率。目前利用CNN（卷积神经网络）和RNN（递归神经网络）技术，百度成功地实现将图像内容生成自然语言的描述性句子或段落，从而在高层语义层面建立了图像和自然语言之间的桥梁，也就是“机器读图”，这可以说是人工智能领域的一次技术飞跃。

3.2 知识图谱

当用户使用搜索引擎时，需要的不止是索引到相关的网页，更希望找到答案、加深了解以及发现更多的内容。为了使搜索引擎更智能，信息的组织方式正在由网页之间的超链联系向海量实体之间的知识联系演变，知识图谱就是基于海量的互联网数据，实现这种演变的最为重要的技术之一。

知识图谱包含了万物以及它们之间的联系，用实体以及实体关系刻画这个世界。如图2所示，百度知识图谱依托于强大的互联网数据分析技术，对互联网海量数据进行挖掘，并应用高效精准的算法对数据进行分类梳理，将复杂的知识体系通过数据挖掘、信息处理、知识计量和图形绘制显示出来，构建宏大的知识网络，以图文

图2 百度知识图谱示例

并茂的方式展现知识的方方面面，让人们更便捷地获取信息、找到所求，这恰恰与百度的使命一脉相承。

为了使互联网中海量的数据及内容为机器所理解，进而形成知识供用户获取并使用，百度知识图谱以实体为基点，创建了基于语义的链接关系，从海量的数据中提取出精华信息，完成了知识的汇集、整理、再加工，构建了与国际标准接轨的数据“智囊”，目前已建成涵盖近20领域、几十类别、上亿实体量的庞大知识数据库。通过强大的平台与灵活的机制，应用到20多个产品线之中，为用户带来多角度、全方位的搜索体验提升。

3.3 深度问答

深度问答是一种基于海量互联网数据和深度语义理解的智能系统，基于对用户自然语言的理解，实现对海量数据的深层分析和语义理解，并通过搜索和语义匹配技术，提炼出答案信息，对信息进行聚合、提炼，给出最全面、准确的结果。其实现的难点主要在于正确理解用户复杂和多变的需求，并掌握海量结构化的知识库数据，这就需要强大的人工智能技术和海量复杂的大数据处理能力。深度问答其关键技术包括问题分析和理解技术、实体知识体系建模技术、文本分析和关系抽取技术以及语义分析和排序技术等。

- 问题分析和理解技术：针对不同类型的问题，提取答案的技术也会不同。根据可采用的技术，问题可以大致分为实体类问题和非实体类问题两大类。实体类问题是指答案是实体的问题，对于实体类问题，问题的答案可以是唯一实体或者实体的列表，需要通过问题分析技术分析出实体类别；对于非实体类的问题，需要通过问题分析技术，把这些类型的问题跟实体类问题区分开来，因为这些问题的答案不再是实体，答案的形态也更加复杂。
- 实体知识体系建模技术：实体类问答离不开实体知识体系的支撑，实体的类别、实体间的同位、上下位关系都十分重要。因此，一个完备的实体知识体系建设（ontology）对于问题回答十分必要。实体的同位、上下位关系可以通过整合多种来源的知识获取，包括一些结构化的数据如百度百科，也可以从普通文本中挖掘。
- 文本分析和关系抽取技术：对文本的深层分析是深度问答用到的一项基础技术。如图3所示，文本的分析分为多个层

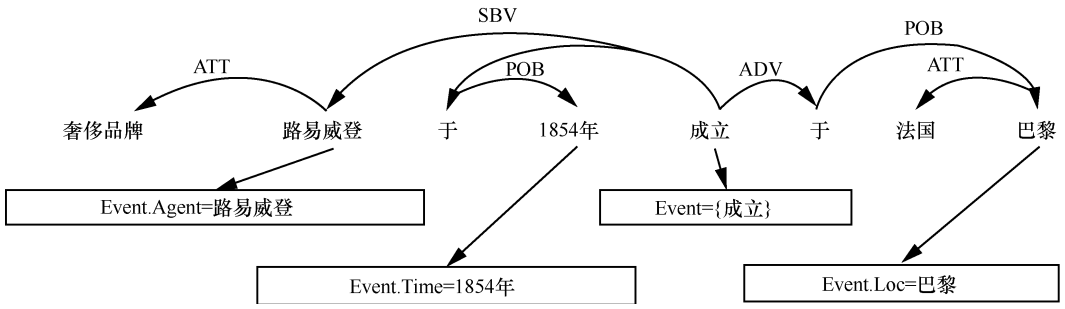


图3 文本分析和知识抽取技术示例

次，包括分词、实体识别、句法分析乃至语义角色标注，在这些分析的基础上可以进行知识获取。而通过对海量数据进行深层分析，可以有效过滤文本分析引入的噪音，使得知识更加精准。文本分析和关系抽取技术不仅可以用于从普通文本抽取知识，也可以用于语义匹配。

4 百度大数据引擎及行业应用实践

4.1 百度大数据引擎

百度坚信技术改变互联网，互联网可以改造传统行业。为了助力传统行业快速进入这个大数据的时代，充分发掘和利用大数据的价值，百度对外发布大数据引擎，向外界提供大数据存储、分析及挖掘的技术能力，这也是全球首个开放大数据引擎。

如图4所示，百度大数据引擎主要包含三大组件：开放云、数据工厂和百度大脑。开放云可以将企业原本价值密度低、结构多样的小数据汇聚成可虚拟化、可检索的大数据，解决数据存储和计算瓶颈；数据工厂对这些数据加工、处理、检索，把数据关联起来，从中挖掘出一定的价值；百度大脑是建立在百度深度学习和大规模机器学习基础上，最终实现更具前瞻性的智能数据分析及预测功能，以实现数据智能，支持科学决策与创造。百度积极开放输出百度大脑的能力，一方面助力国家在人工智

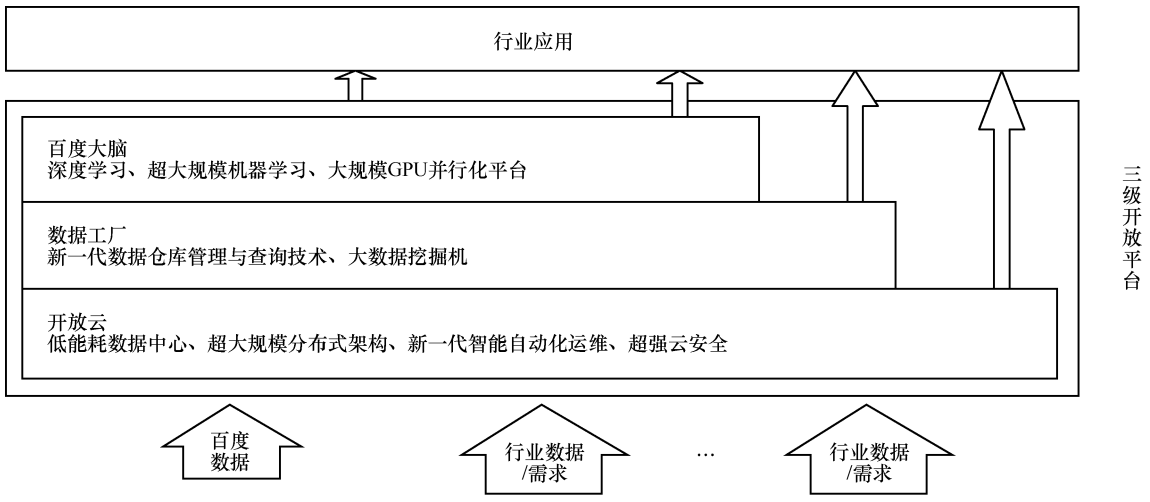


图4 百度大数据引擎

能、大数据等技术上的整体提升；另一方面也帮助行业转型升级，提升企业的核心竞争力。

这三大组件作为3级开放平台支撑百度核心业务及其拓展业务，也将作为独立或整体的开放平台，给各行各业提供支持和服务，支持百度的核心商业应用及社会企业的新兴商业模式。

4.2 百度行业应用大数据实践

4.2.1 公众生活领域——大数据预测

百度基于海量的数据处理能力，利用机器学习和深度学习等手段建立模型，可以实现公众生活的预测业务。目前，在百度预测产品中已经推出了景点舒适度预测和城市旅游预测、高考预测、世界杯预测等服务。

以世界杯预测为例，在2014年巴西世界杯的四分之一决赛前，百度、谷歌、微软和高盛分别对4强结果进行了预测，结果显示：百度、微软结果预测完全正确，而谷歌则预测正确3支晋级球队；在小组赛阶段的预测，谷歌缺席，微软、高盛的准确率也低于百度[6]。总体来看，无论是小组赛还是淘汰赛，百度的世界杯结果预测中均领先于其他公司。最终，百度又成功预测了德国队夺冠，如图5所示。

预测准确度来自百度对大数据的强大分析能力和超大规模机器学习模型。在对体育数据的研究过程中，百度的科学家发现类似保罗章鱼的赛事预测完全有可能借助大数据的分析能力完成。因此，百度收集了2010-2013年全世界范围内所有国家队及俱乐部的赛事数据，构建了赛事预测模型，并通过对多源异构数据的综合分析，综合考虑球队实力、近期状态、主场效应、博彩数据和大赛能力等5个维度的数据。最终实现了对2014年巴西世界杯的成功预测。

4.2.2 公共卫生领域——疾病预测

通过百度搜索数据与医疗数据、医保数据等关联，并结合图像识别和语音识别技术、可穿戴设备数据采集等，通过大数据分析与挖掘能力可以实现人群疾病分布关联分析等。通过对大量临床电子病历、临床经验和科研成果等医学信息数据进行学习和理解，绘制人类疾病图谱（人群分布），并建立疾病分析模型和治疗路径模型。这也将极大推动疾病研究、医药研发、药品监管、居民医疗服务和全民健康教育等事业发展。

百度与中国疾病预防控制中心（CDC）合作开发的疾病预测产品，基于对网民每日更新的互联网搜索的分析、建模，实时反馈流感、手足口、性病、艾滋病等传染病，糖尿病、高血压、肺癌、乳腺癌等流行病的爆发数据，并预测疾病流行趋势，是国家疾病控制机构传统监测体系的有力补充。结合大数据舆情分析、公共卫生危机事件预警产品，有效地融合非结构化大数据，建立了基于互联网的新兴公共卫生数据资源共享机制与服务价值链。

6 http://www.ithome.com/html/it/93409.htm

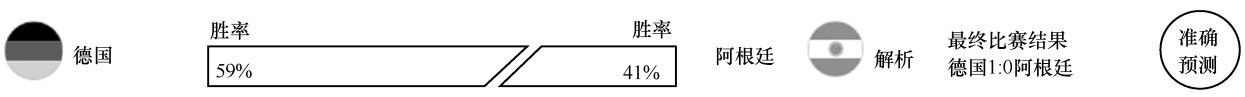


图5 百度世界杯预测

4.2.3 企业IT应用——硬盘故障预测

百度全球有几十个的数据中心或者内容分发网络（CDN）节点，拥有数十万台服务器和数万台交换机，200多万块硬盘。这些硬盘的年报错率为4%~7%，月均硬盘故障超过1万起，占全部硬件故障的80%以上。百度通过大数据分析与机器学习技术，对9亿条实例进行采集处理，选取15万个训练样本，监控240个特征的实时变化，构建预测模型，并通过机器学习的算法可以提前一天预测出硬盘故障并迁移数据，该系统可以节约带宽70%、节约计算资源85%、节省服务器运行消耗10%，每年节省1万多块硬盘。如图6所示，基于大数据实现硬盘故障预测的方法也可以用于实现行业硬件系统的运维和管理中。

4.2.4 企业IT应用——智能化运维

近年来百度在服务器规模、数据规模、单集群规模等方面出现爆发式增长。百度服务器的规模近5年来增长了15倍以上，达到数十万台。数据规模已达到EB级别。在云计算和大数据时代，集群规模和数据量爆发式增长，如何管理好云计算平台、如何提供高质量的服务，是云计算的核心问题之一。

为了应对云计算和大数据应用带来的新的需求和挑战，百度同样利用大数据技术，把在线服务运维转向智能化管理模式，并走在了行业的前列。百度已经建立起了六大数据仓库之一的运维数据仓库，囊括了服务器、网络、系统、程序、变更等各个方面的实时及历史状态数据，每天更新

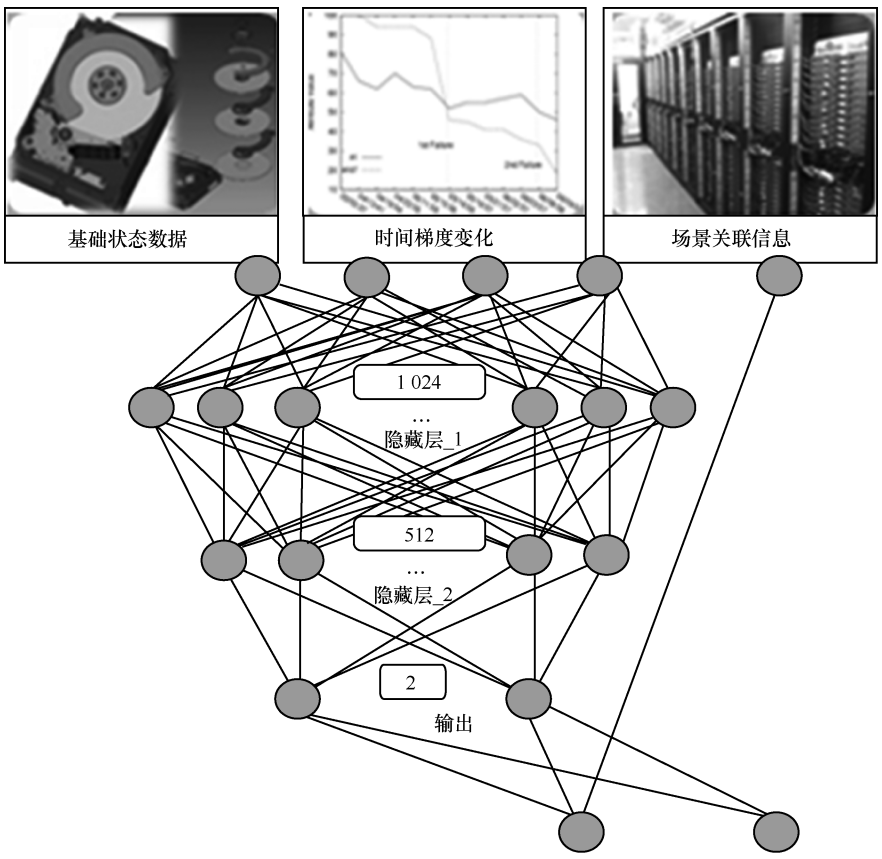


图6 基于大数据的硬盘故障预测

数据量接近100 TB。

基于对运维大数据的挖掘、对历史数据的学习和异常模式识别，实现对流量数据的预测。通过对包括访问速度、系统容量、带宽、成本等在内的10多个因子的实时自动分析，实现了在众多数据中心间的流量自动调度，决策时间也由人工判断的10几分钟大幅缩短到1 min。这个系统的实际效果在故障中得到很好的检验，例如系统在没有人工介入的情况下智能地把流量调度到另外的数据中心，拒绝流量仅有几千个，避免类似故障可能造成数千万的流量损失。

4.2.5 社会治理领域——上海外滩踩踏事故大数据分析

用户去目的地之前，一般都会提前利用百度地图搜索地点和规划路线。同时，百度的搜索词也会有一定的提前量预测某一事件。因此，对百度数据的分析可以应用于社会治理领域，实现基于大规模人群的事件预警和分析。

2015年初的上海外滩踩踏事件发生后，百度秉承“以数据说话”的理念，通过对百度的定位数据、搜索数据进行挖掘，对当时的情况进行了数据化描述。图7标明了南京东路地铁站附近区域、外滩源附近区域、事发地陈毅广场附近区域和外滩区域位置在2014年12月31日事发当时的人群热力图。颜色越深表示人群越密集，颜色越浅表示越稀疏。

对当晚外滩区域的人流进行量化分析，得到了如图8所示的人群流动方向分布情况。图8中每一扇形分区代表不同的人流方向，扇区半径表示该方向人流量大小。图8（a）和图8（b）表示2014年中秋和国庆当晚的情况，可以看出，人流方向比较简单和清晰，即南北向人流较多，其他方向人流较少。图8（c）显示了跨年当晚外滩区域的人流方向，除了南北双向的人流，还有其他多个方向人流，人群流动方向分布混乱。

为了挖掘用户行为的时空特性，百度对大量历史群体聚集场合的数据进行进一步分析，包括鸟巢足球赛等。分析发现，

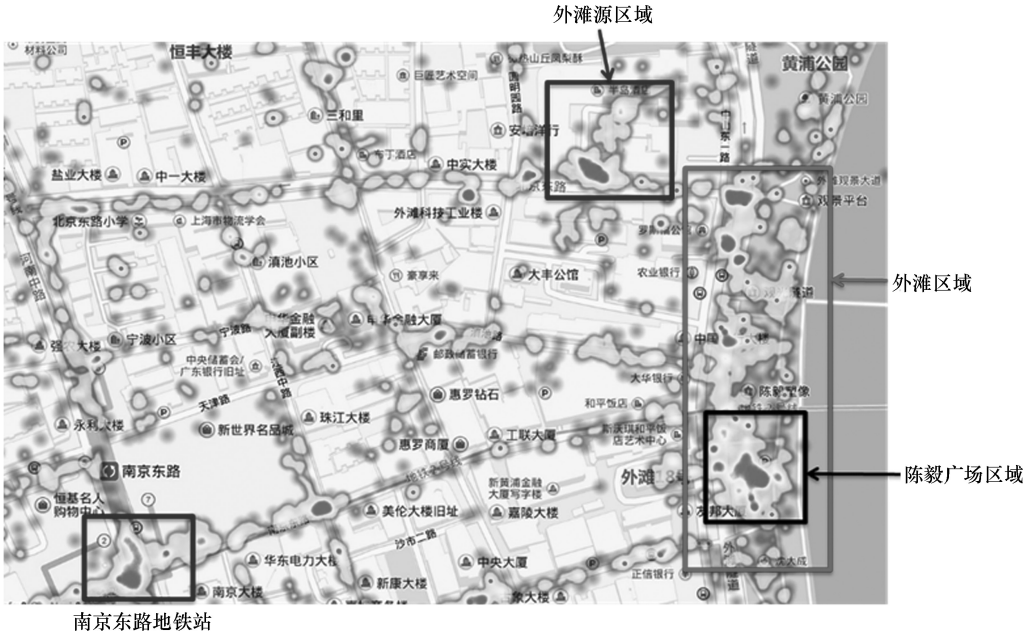
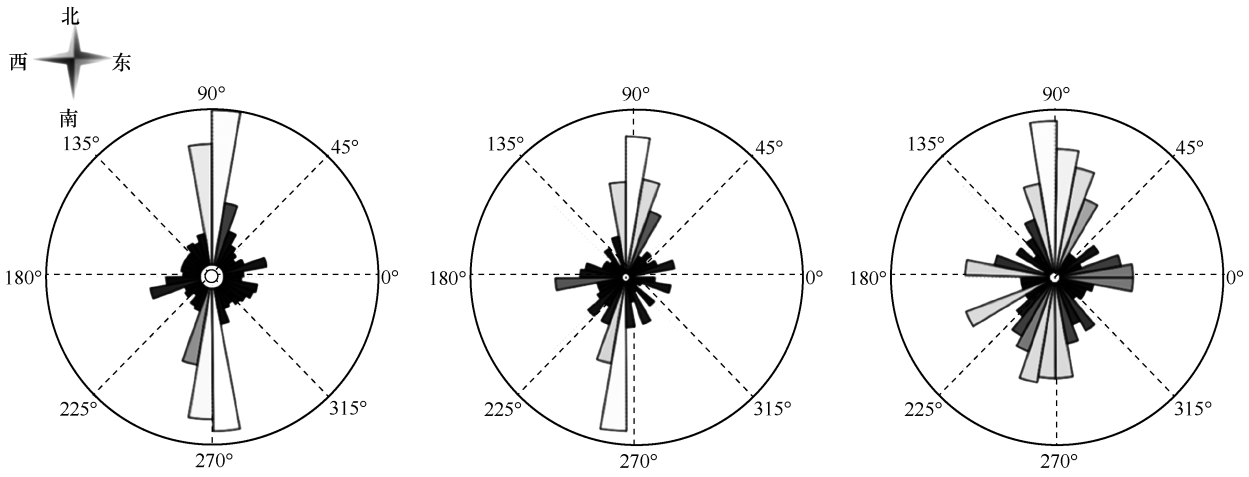


图7 外滩地区人群热力图



(a)2014年中秋前夜　(b)2014年国庆当晚　(c)2014年跨年当晚

图8　人群流动方向分布情况

相关地点的地图搜索请求峰值会早于人群密度高峰几十分钟出现。图9为外滩的搜索量和人群数量之间的互相关性相对于时延的变化曲线，其中横轴的值为时延量，负值表示提前量。例如，横坐标-10对应的纵坐标值就是提前10 h的搜索量与人群数量的相关性。从图9中可以发现，两个量的互相关性曲线在-1.5 h时达到了峰值，这意味着，根据地图上相关地点搜索的请求量，至少可能提前几十分钟预测出人流量峰值的到来。

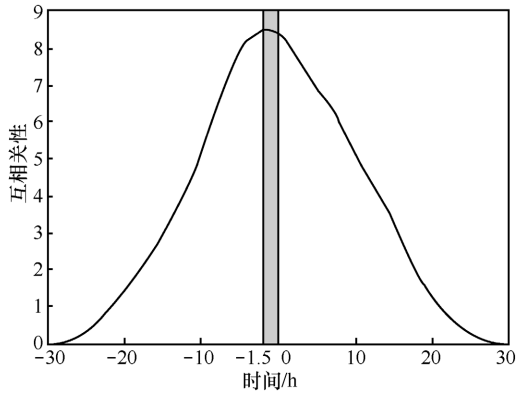


图9　搜索量和人群数量相关性曲线

5 结束语

随着我国各行业信息化的快速发展，数据量激增，我国已经成为数据大国。未来如何将这些数据得以有效、科学地利用，挖掘数据价值，将我国建设为大数据技术强国，是信息化发展的重要战略问题。进入大数据时代，数据类型已不是单一的结构化数据，非结构化数据占有非常大的比重，但是如果现有技术手段无法将大量的非结构化数据与结构化数据进行统一和整合，就无法发掘数据中的重要价值。而对于这些非结构化的数据进行分析和挖掘并实现其价值，人工智能是重要的技术发展方向。大数据和计算技术的发展带来了人工智能的新浪潮，人工智能的本质特征之一是学习的能力，也就是说系统的性能会随着经验数据的积累而不断提升。所以，大数据时代的到来给人工智能的发展提供前所未有的机遇。

如图10所示，在人工智能领域，存在着一个正循环：通过人工智能技术不断优化产品，让优秀产品吸引更多用户，更多用户产生

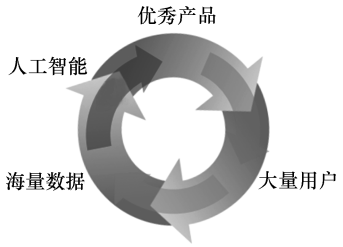


图10 基于大数据的人工智能正循环

更多数据，而更多的数据可以使人工智能的性能得到提升，从而让产品更优秀。

在过去的20年里，中国企业很多时候都只能扮演技术跟随者的角色，但是现阶段我国互联网企业在大数据处理和人工智能等领域不断取得突破，推动了这个正循环运转加速，引领我国信息技术的发展，并在世界范围内树立技术强国的形象，推动我国的大数据产业成熟和发展。

参考文献

[1] 涂兰敬. 百度的技术突破与应用. 中国计算机报, 2015-01-05
Tu L J. Technology breakthrough and application of the Baidu. Chinese Computer Newspaper, 2015-01-05

[2] 都大龙, 余轶男, 罗恒等. 基于深度学习的图像识别进展：百度的若干实践. 中国计算机学会通讯, 2015, 11(4)
Du D L, Yu Y N, Luo H, *et al*. Progress of image recognition based on deep learning:some of the Baidu practice. Communications of the CCF, 2015, 11(4)

作者简介

陈尚义，百度技术委员会理事长，国家科技重大专项（03专项）总体组专家，中国电子学会常务理事，中国电子学会、中国计算机学会大数据专家委员会委员，北京航空航天大学、合肥工业大学兼职教授，北京航空航天大学计算机校友会会长。先后就职于国家发展改革委员会办公厅、国家开发银行科技局从事信息化工作，新加坡国立大学、美国硅谷高科技公司从事信息技术产品的研发工作，2011年初加入百度。获省部级科技进步奖一等奖1次，二等奖3次、三等奖4次，2009年度“北京市创新人物”。

收稿日期：2015-05-04；**修回日期**：2015-05-06
论文引用格式：陈尚义. 百度大数据应用与实践. 大数据, 2015009
Chen S Y. Big data applications and practices of Baidu. Big Data Research, 2015009